BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e37048
doi: 10.3897/biss.3.37048

OPEN ACCESS

Conference Abstract

A Machine Learning Based Approach for Similarity Search on Biodiversity Knowledge Graphs

Claus Weiland[‡], Maxat Kulmanov[§], Marco Schmidt[‡,|], Robert Hoehndorf[§]

‡ Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany
§ King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
| Palmengarten der Stadt Frankfurt am Main, Frankfurt am Main, Germany

Corresponding author: Claus Weiland (cweiland@senckenberg.de)

Received: 10 Jun 2019 | Published: 13 Jun 2019

Citation: Weiland C, Kulmanov M, Schmidt M, Hoehndorf R (2019) A Machine Learning Based Approach for Similarity Search on Biodiversity Knowledge Graphs. Biodiversity Information Science and Standards 3: e37048. https://doi.org/10.3897/biss.3.37048

Abstract

Mass biodiversity data from scientific collections will be provided by world-wide digitization efforts like iDigBio in the U.S and DiSSCo in Europe. This opens up an increasing amount of data on wild type organisms, which enables the building of large biodiversity knowledge graphs comprising, *inter alia*, sequence, trait and occurrence data. Knowledge graphs model information in the form of entities and their relationships expressed in good practice as ontology-based annotations. Based on ontological descriptions, semantic similarity analysis makes linking of wild type data to genomic and proteonomic data of model organisms possible and thus supports knowledge discovery of crop wild relatives and underutilized species of interest for medicine, breeding and agriculture. Since classical similarity measurements focus on recording differences between character states (aiming to describe disease phenotypes), but not the character states in the sense of trait variations itself, new methods for similarity search are required. Machine learning algorithms operate on feature vectors, which are numeric representations of data (images, class labels etc) in n-dimensional vector space. We established a machine learning based workflow for similarity search on biodiversity entities using feature learning on ontologies and an associated RDF knowledge graph to project structured trait data into vector space. Vectors are then compared applying a similarity function (e.g. cosine similarity) to determine similarity between taxa based on trait semantics. We will present an

application example of machine learning on biodiversity knowledge graphs using a pipeline built upon OPA2Vec, a method to generate feature vectors from the logical content of ontologies (Smaili et al. 2018), to successfully cluster plant species for life form and ecotype (e.g. tree vs. perennial plant) on the basis of their annotations with the Flora Phenotype Ontology (Hoehndorf et al. 2016).

Keywords

semantic similarity, machine learning, trait semantics, phenotype ontology, knowledge graph

Presenting author

Claus Weiland

Presented at

Biodiversity_Next 2019

References

- Hoehndorf R, Alshahrani M, Gkoutos GV, Gosline G, Groom Q, Hamann T, Kattge J, de Oliveira SM, Schmidt M, Sierra S, Smets E, Vos RA, Weiland C (2016) The flora phenotype ontology (FLOPO): tool for integrating morphological traits and phenotypes of vascular plants. Journal of Biomedical Semantics 7 (1): 65. https://doi.org/10.1186/s13326-016-0107-8
- Smaili FZ, Gao X, Hoehndorf R (2018) OPA2Vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction. Bioinformatics bty933. https://doi.org/10.1093/bioinformatics/bty933